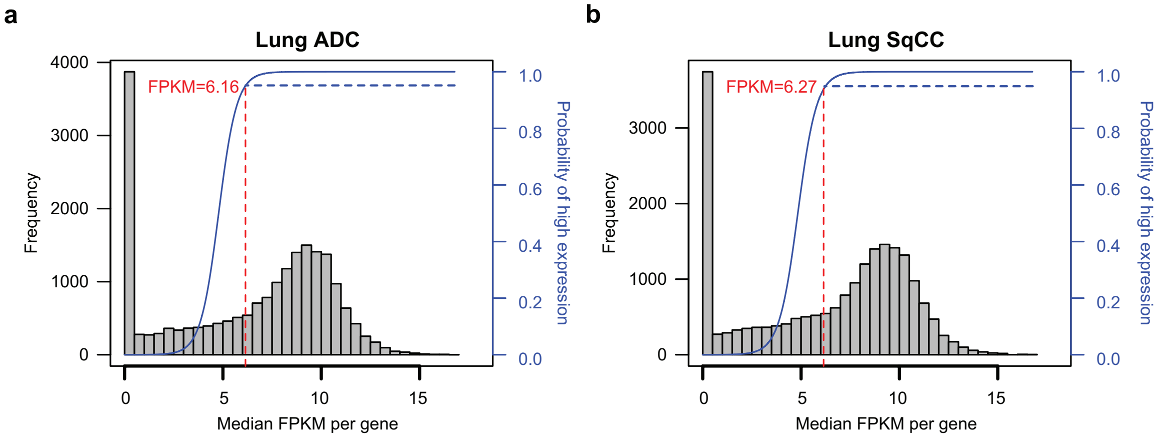
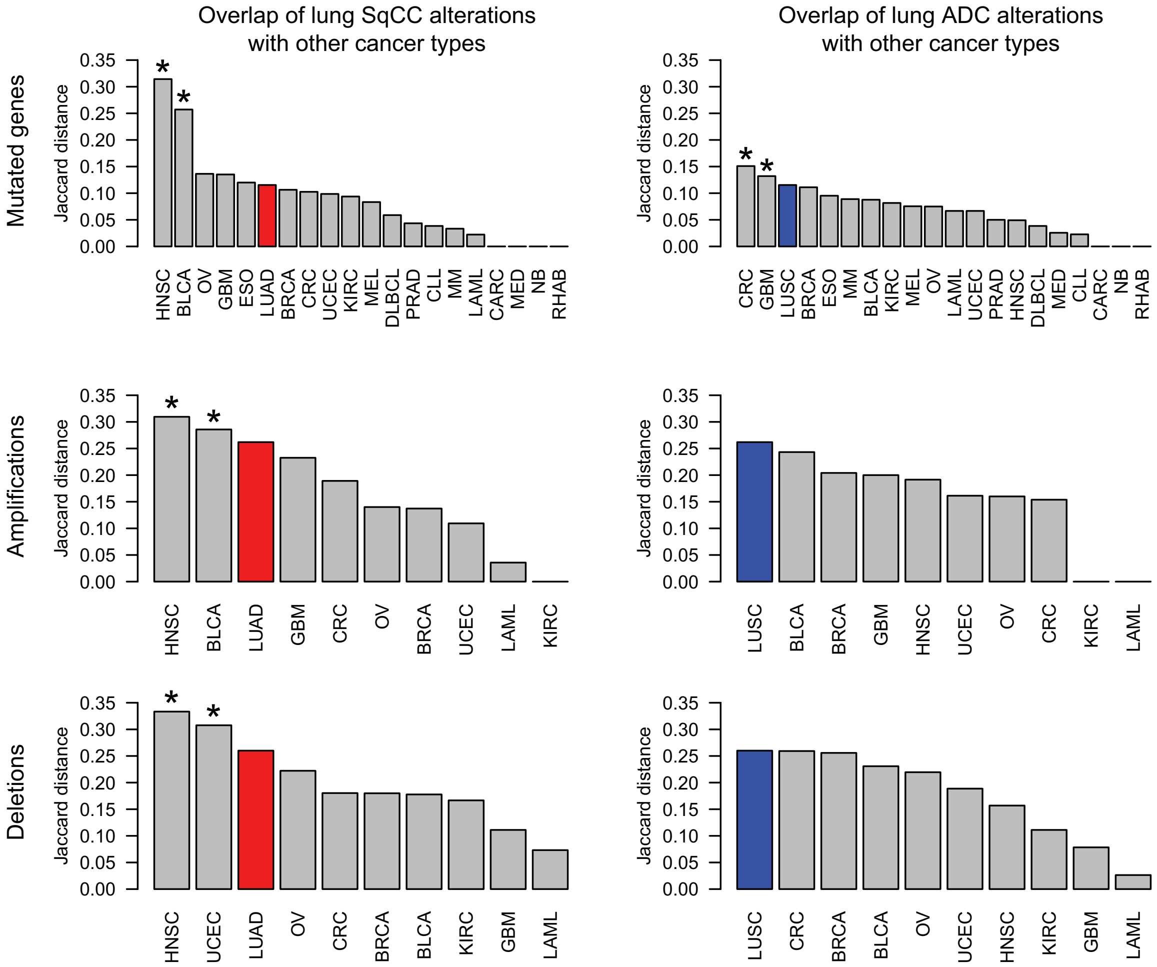
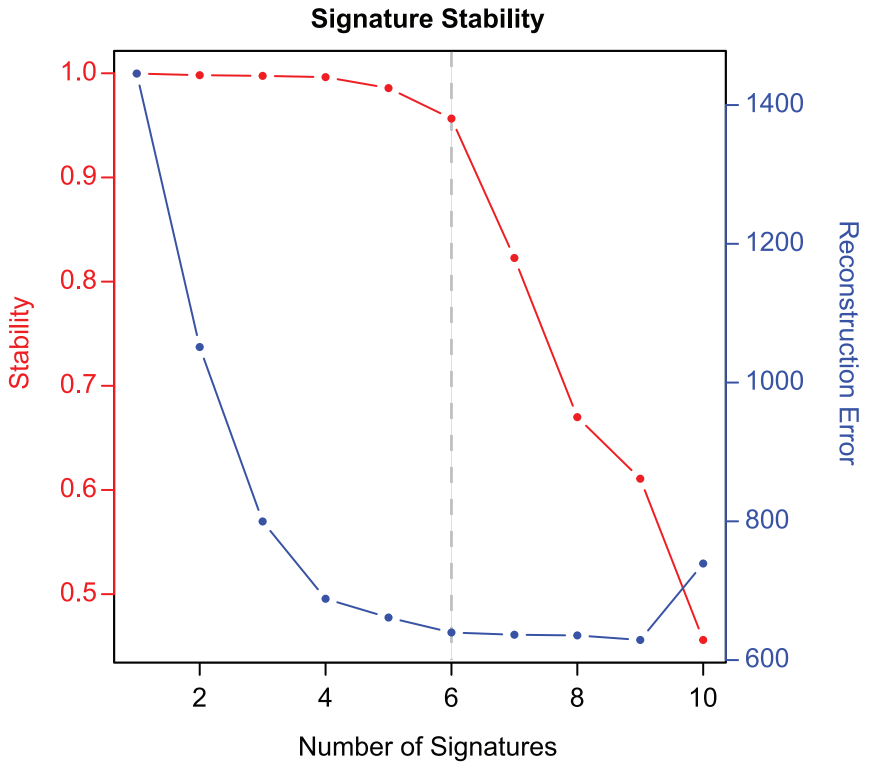


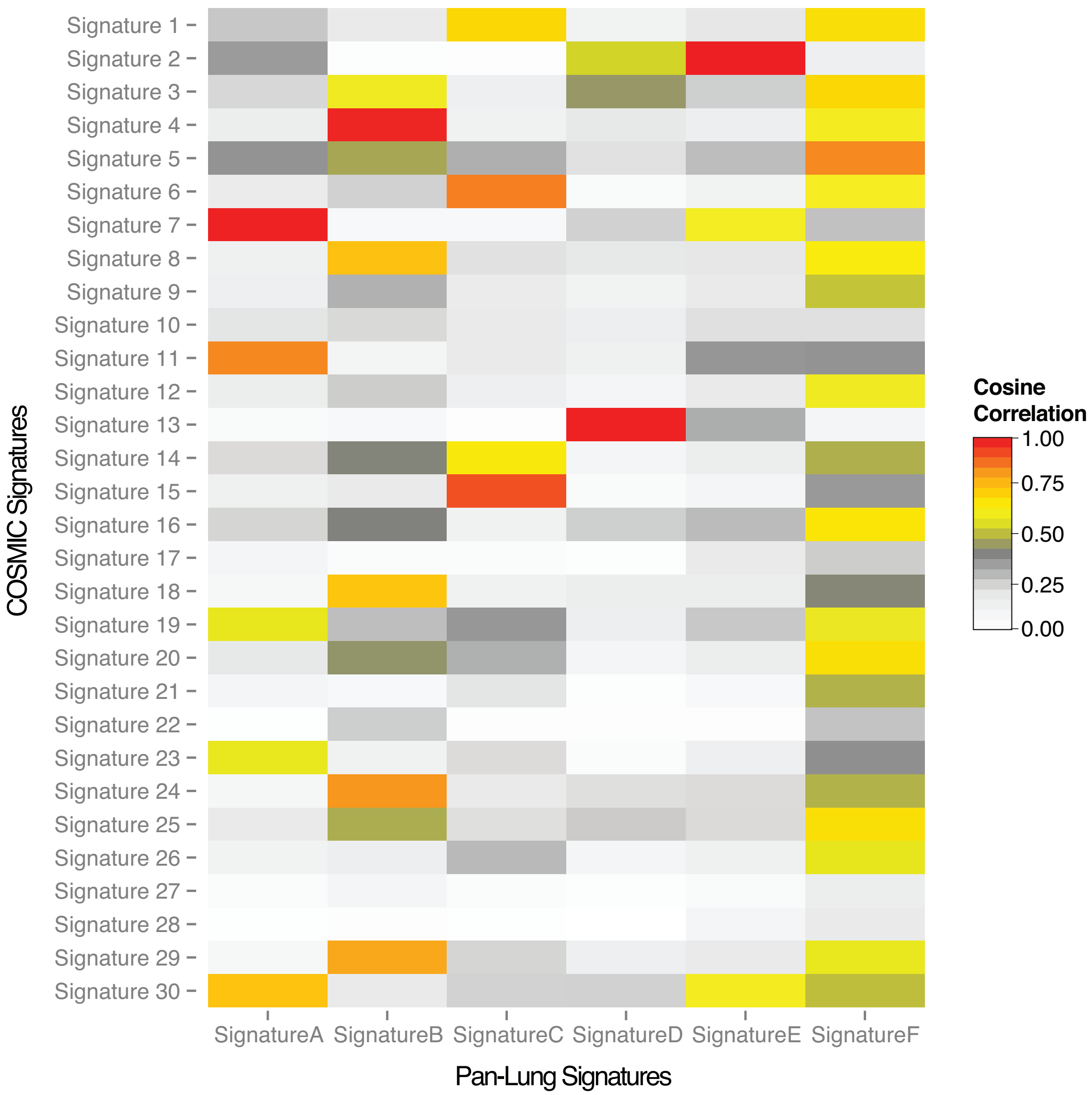
Supplementary Figure 1. Distribution of median FPKM expression values for relatively pure lung ADCs and SqCCs. Using **(a)** ADC tumors or **(b)** SqCC tumors that had >50% ABSOLUTE-estimated purity and a matching RNA-seq sample, a mixture model of two normal distributions was fit using the R Mclust package using default parameters. Only genes that had a >95% probability of belonging to the cluster with higher expression were considered in the final FDR correction in the MutSig2CV analysis. This threshold corresponded to 6.16 $\log_2$ FPKM for ADCs and 6.27 $\log_2$ FPKM for SqCCs as indicated by the dotted line.



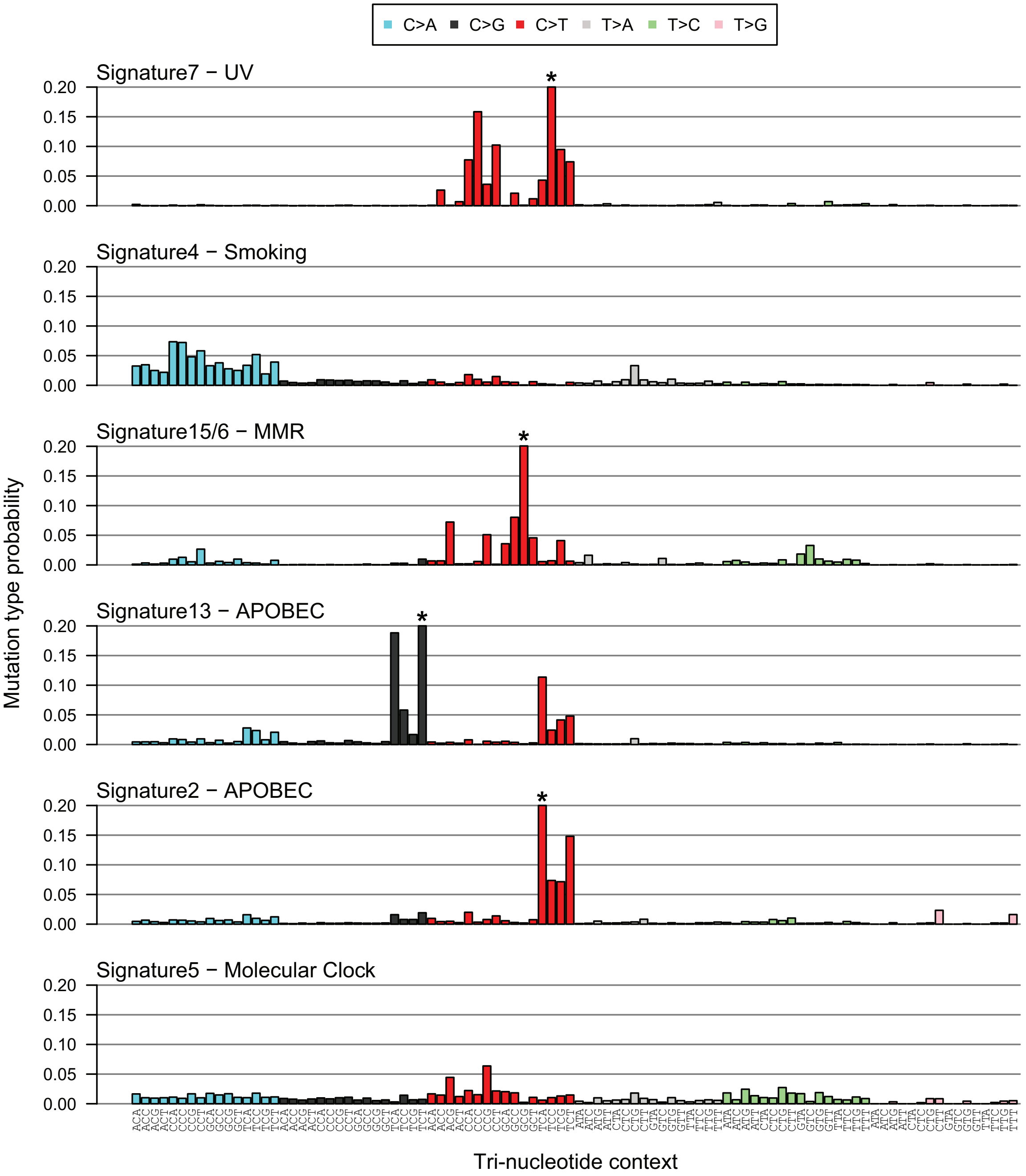
Supplementary Figure 2. Comparison of somatic alterations between lung ADC, lung SqCC, and other tumor types. The overlap between lists of genes that were significantly mutated, focally amplified, or focally deleted for lung ADCs (abbreviated as LUAD in the barplots) and lung SqCCs (abbreviated as LUSC in the barplots) and lists of genes from other cancer types was examined using the Jaccard index. The Jaccard index is the number of genes in the intersection of two gene lists divided by the number of genes in the union of two gene lists. Thus, a higher Jiccard index indicates a higher degree of similarity between two gene lists. Additionally, we used the Fisher's exact test to determine if the overlap between two gene lists is significant given the background set of all significantly mutated genes (or focal copy number peaks) from all tumor types. For lung SqCC, head and neck squamous cell (HNSC) carcinomas had a high degree of overlap in significantly mutated genes, focal amplifications, and focal deletions, while bladder urothelial (BLCA) carcinomas had a high overlap in both significantly mutated genes and amplifications. In contrast, the significantly mutated genes in lung ADC were most similar to glioblastoma (GBM) and colorectal cancer (CRC) albeit to a less extent when comparing lung SqCC to HNSC and BLCA. While lung ADC and lung SqCC did share a large number of focal deletion peaks (n=13), five of these peaks were putative fragile sites. An asterisk indicates a FDR *q*-value < 0.10.



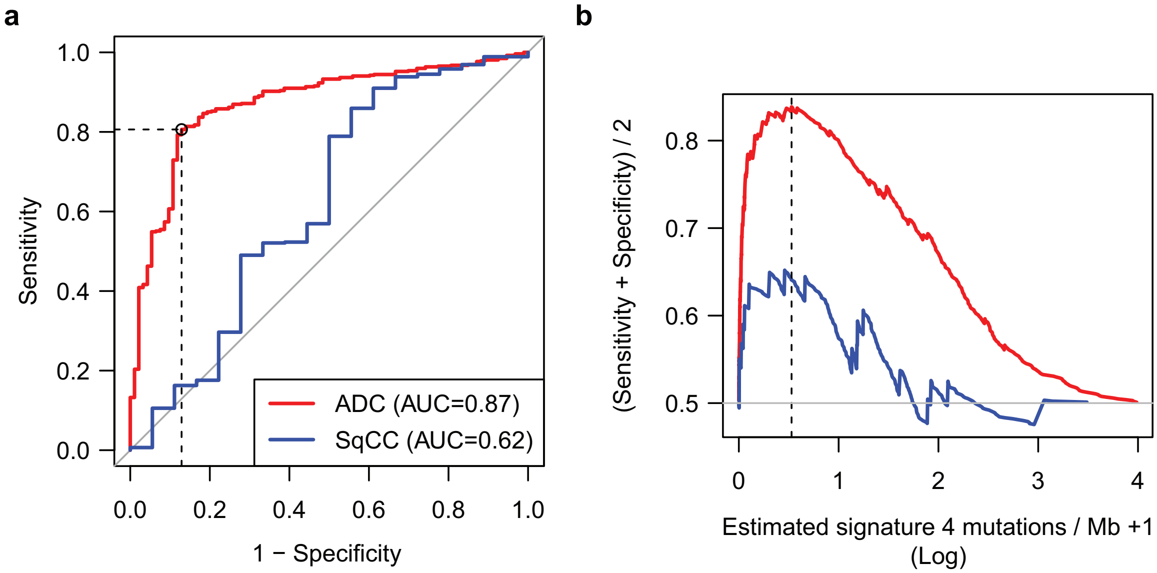
Supplementary Figure 3. Mutational signature identification using non-negative matrix factorization (NMF). NMF was run with 1,144 lung cancer exomes using 6 mutation types with 16 different trinucleotide contexts and 2 transcriptional strands (i.e. is the mutation on the transcribed strand or not) for a total of 192 mutational states (**Supplementary Table 7**). The number of possible signatures was varied from 1 to 10. The signature stability and the Frobenius reconstruction error were assessed via bootstrapping as previously described[1]. Signature stability evaluates the similarity between the extracted mutational signatures from stochastically initialized iterations and low reconstruction error is indicative of an accurate description of the original cancer genome mutation counts. We used the solution with 6 signatures as the stability dropped below 0.90 with the expansion to additional signatures.



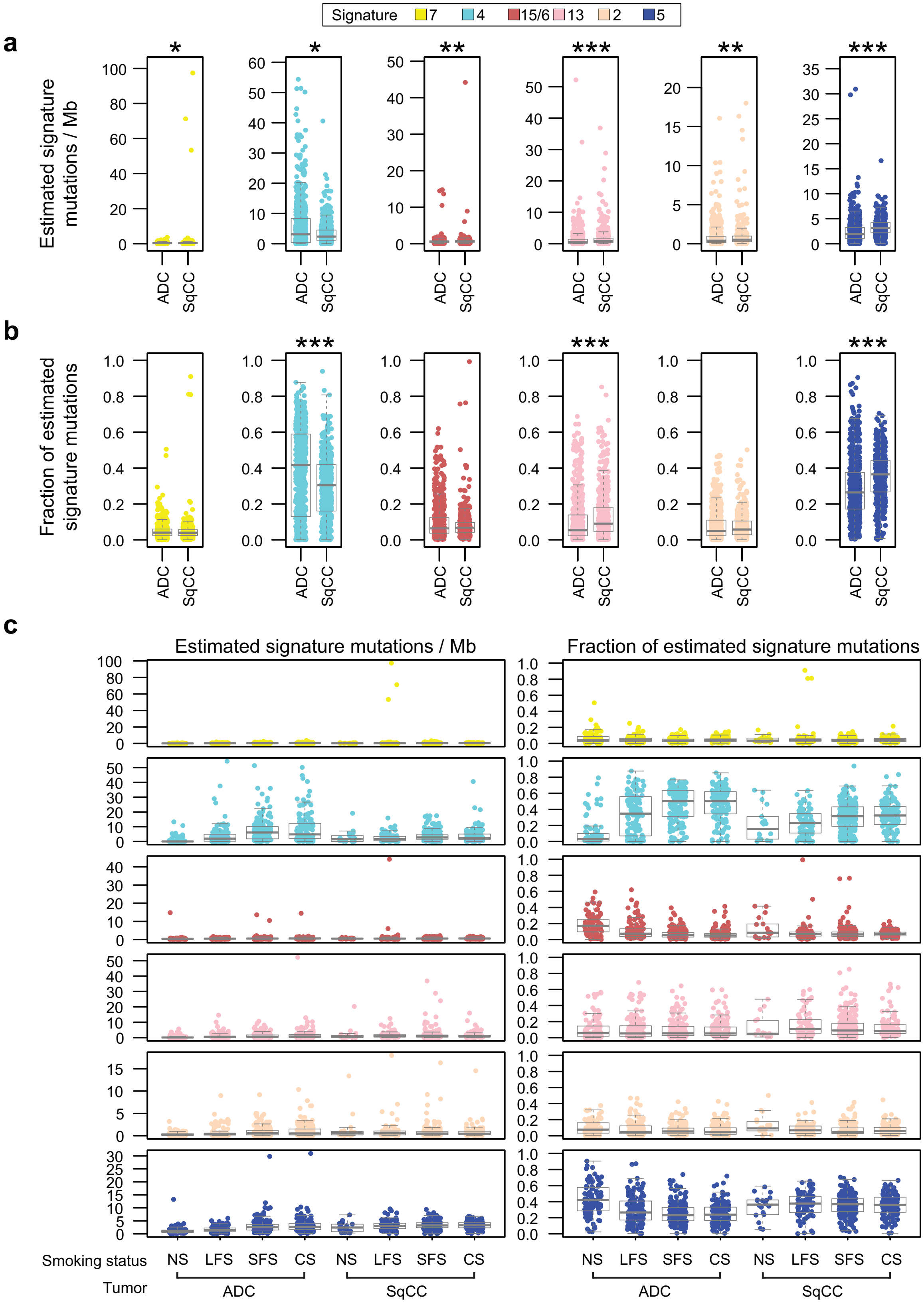
Supplementary Figure 4. Correlation between mutational signatures derived in the Pan-Lung cohort and previously defined signatures from COSMIC. The trinucleotide probabilities for 30 signatures generated from 10,952 exomes and 1,048 whole-genomes across 40 were obtained from COSMIC (http://cancer.sanger.ac.uk/cancergenome/assets/signatures_probabilities.txt). A pair-wise cosine correlation was performed between all COSMIC and Pan-Lung signatures. The top one or two most correlated COSMIC signatures were used determine the identity of each Pan-Lung signature and each Pan-Lung signature was renamed accordingly.



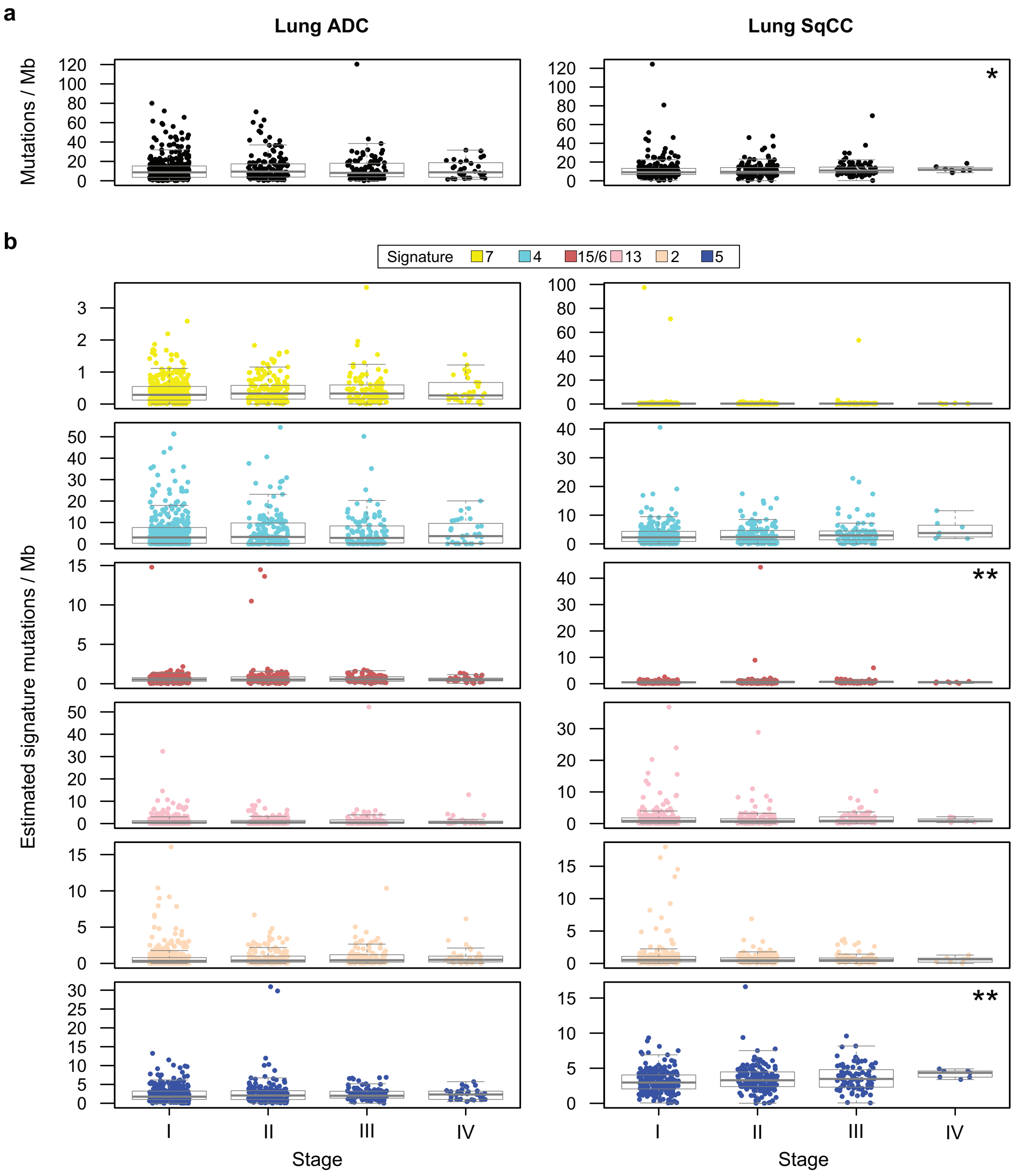
Supplementary Figure 5. Trinucleotide context for six inferred mutational signatures. Mutational signatures included a UV-related signature of C>T at TpCpC/CpCpC, a smoking-related signature of C>A transversions, a mismatch repair (MMR) signature of C>T at GpCpG, APOBEC-related signatures of C>T or C>G at TpCpT/TpCpA, and a moderate enrichment of C>T at CpGs. Asterisks indicate that the mutation type probability exceeds 0.20.



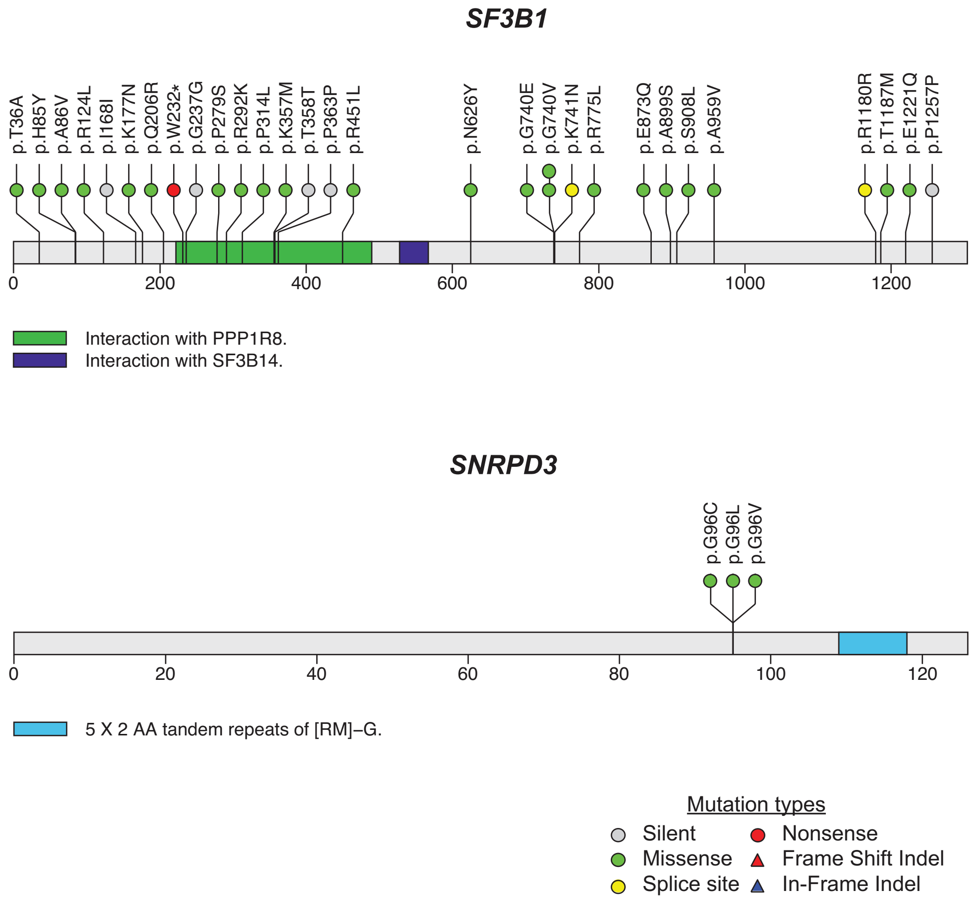
Supplementary Figure 6. Classification of never vs. ever smokers using mutational signatures. (a) The number of estimated SI4 mutations per Mb was used to classify lung ADCs from never smokers (n=93) vs. lung ADCs from ever smokers (n=521) or lung SqCCs from never smokers (n=18) vs. lung SqCCs from ever smokers (n=455). The area under the curve (AUC) was substantially better for lung ADCs (AUC=0.87) compared to lung SqCCs (AUC=0.62). **(b)** The average of the sensitivity and specificity was calculated for each possible threshold of the SI4 mutation rate. The most optimal point in lung ADCs corresponded to a SI4 mutation rate of 0.696 per Mb (sensitivity=0.81; specificity=0.87) as indicated by the dotted lines. Both lung ADCs and lung SqCCs were classified as transversion-low or transversion-high according to this threshold. These results show that the smoking-related mutational signature is not able to distinguish never and ever smokers in lung SqCC and may suggest that the smoking statuses for the 18 never smokers with lung SqCC are inaccurate.


Signature 7 4 15/6 13 2 5
a
Estimated signature mutations / Mb
*
*
**
***
**
***
ADC
SqCC
b
Fraction of estimated signature mutations
***
***
***
c
Estimated signature mutations / Mb
Fraction of estimated signature mutations
Smoking status
NS
LFS
SFS
CS
Tumor
ADC
SqCC

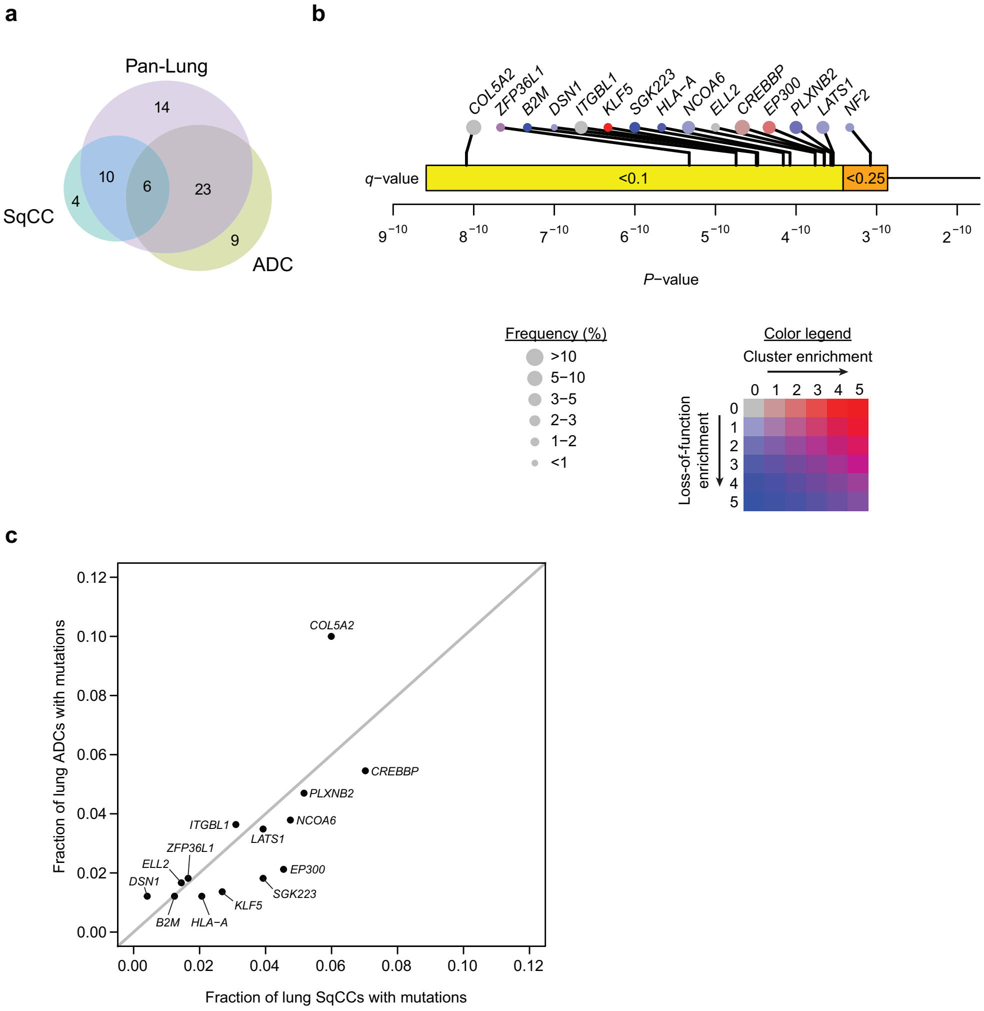
Supplementary Figure 7. Mutational signatures by tumor type and smoking status. (a) The number of estimated mutations per Mb from each signature is shown for each tumor as a function of tumor type. Asterisks indicate significance level from a Wilcoxon rank-sum test (*p < 0.05, **p < 0.01, ***p < 0.001). **(b)** Within each tumor, the fraction of estimated mutations for a signature was derived by dividing the number of estimated mutations for that signature by the sum of estimated mutations from all signatures. This fraction is shown as a function of tumor type. **(c)** The number of estimated signature mutations per Mb within each tumor (left) or the fraction of estimated signature mutations within each tumor (right) is shown for life-long never smokers (NS), longer-term former smokers (LFS) who had quit for at least 15 years, shorter-term former smokers (SFS) who had quit within 15 years, and current smokers (CS) in both lung ADCs and lung SqCCs. Only tumors with available smoking history and duration were included (n = 1001).



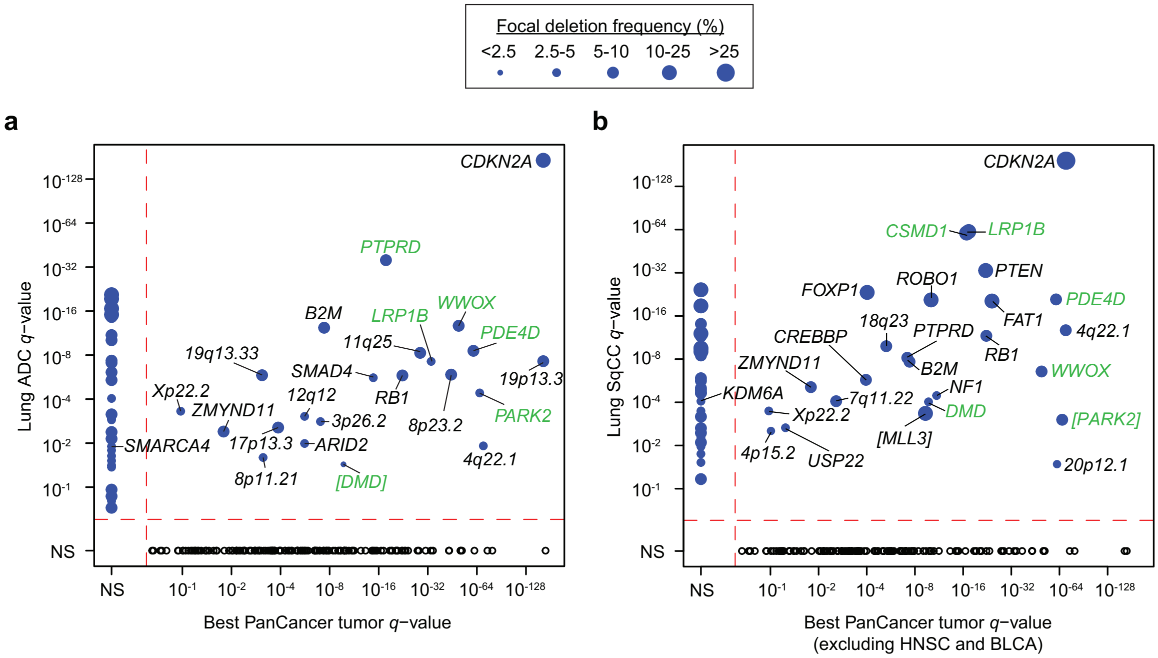
Supplementary Figure 8. Mutational signatures and mutation rates by tumor stage. **(a)** Total mutation rates were associated with stage in lung SqCC (p = 0.036) but not in lung ADC (p > 0.05). **(b)** Estimated rates of mutational signatures per Mb were not associated with stage with the exception of SI15/SI6 and SI5 in lung SqCC (p < 0.01). Asterisks in the top right corner indicate significance level from a Kruskal-Wallis test (*p < 0.05, **p < 0.01, ***p < 0.001).



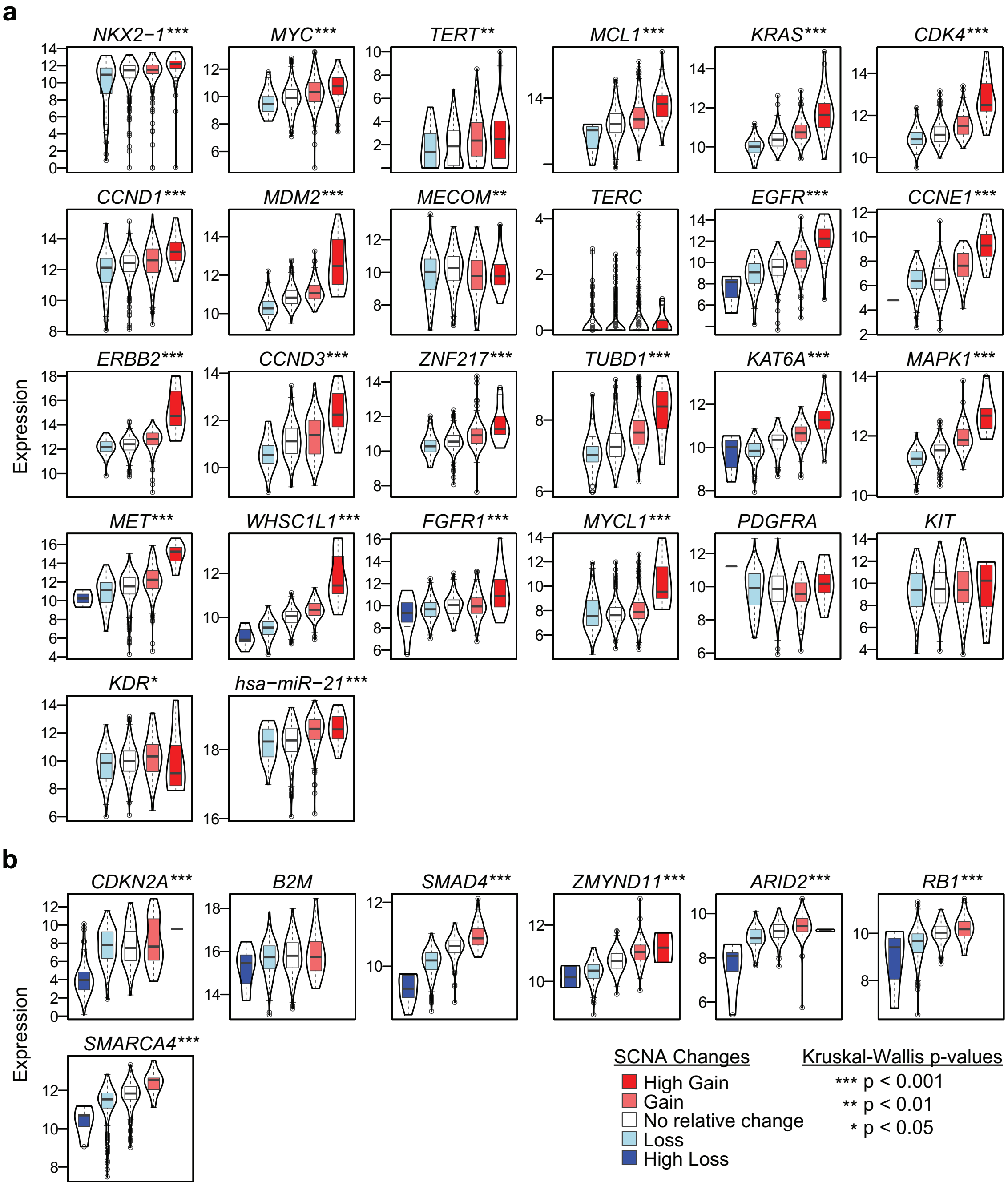
Supplementary Figure 9. Somatic alterations in splicing factors. In addition to *U2AF1*, *RBM10*, and *FTSJD1*, several other splicing factors or RNA binding proteins had recurrent mutations. Recurrent *SF3B1* mutations have been previously observed in chronic lymphocytic leukemia (CLL) in or around amino acid position 740[2]. Additionally, *SF3B1* was significantly mutated in both CLL and breast cancer (BRCA)[3]. Finally, a recurrent mutation was observed in *SNRPD3*, another component of the spliceosome.



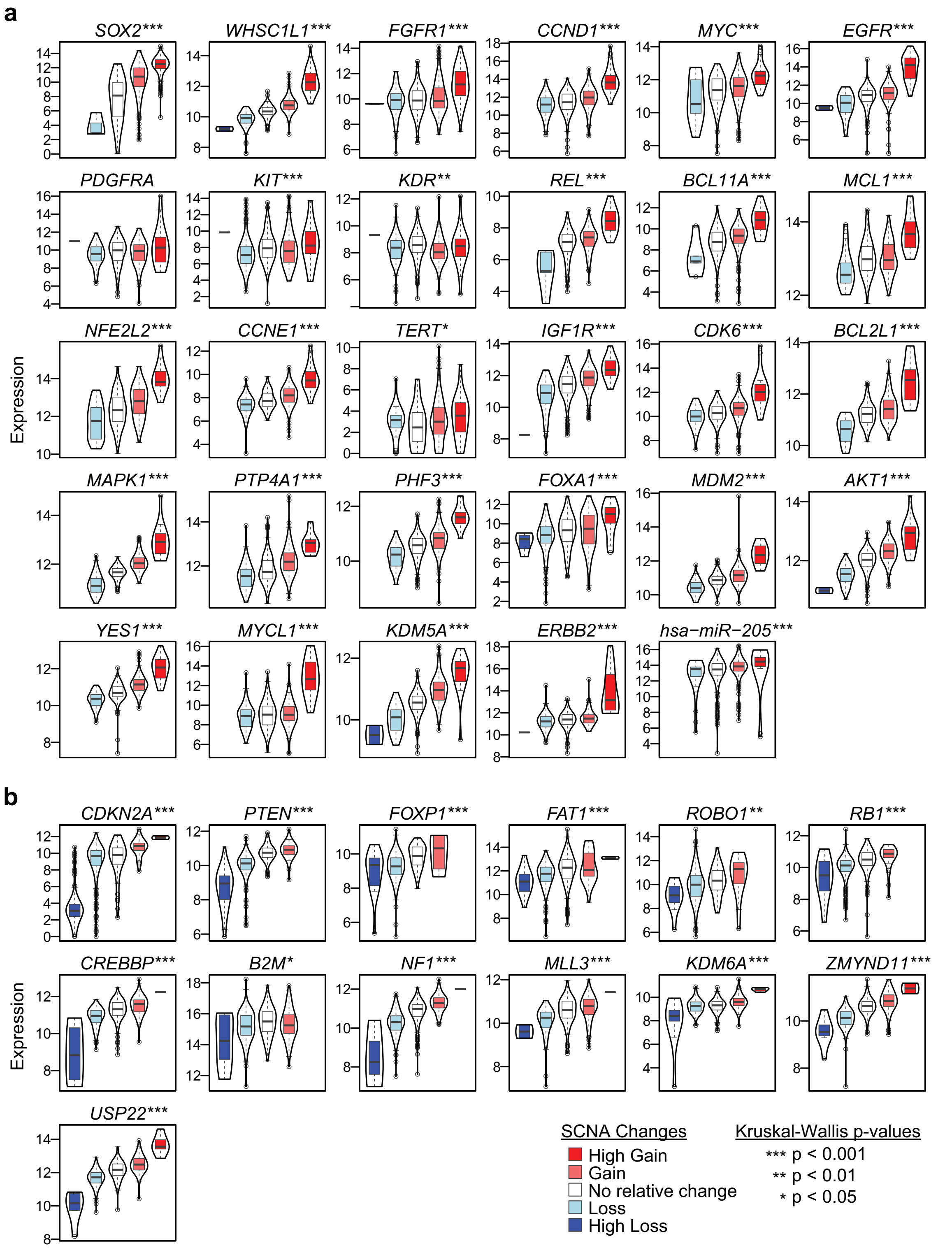
Supplementary Figure 10. Significantly mutated genes identified in the Pan-Lung analysis. Lung ADC and SqCC cohorts were combined and analyzed with MutSig2CV. **(a)** Venn diagram showing the overlap of significantly mutated genes found in each analysis (FDR *q*-value < 0.1). Fourteen additional genes were found to be significantly mutated in the Pan-Lung cohort that were not identified as significantly mutated in either individual tumor type. **(b)**. Genes were ranked from those with the smallest p-value (left) to the largest p-value (right) from the Pan-Lung analysis. The positions of the 14 additional Pan-Lung genes are indicated. Size of the point indicates the frequency of mutations in the gene in the PanLung cohort and the color of the point indicates enrichment for mutation clustering or enrichment for loss-of-function mutations as in **Figure 3**. Previously characterized cancer genes included immune related genes such as *HLA-A* and *B2M* as well as the paralogs *EP300* and *CREBBP*. A novel cancer gene included the transcription factor *KLF5*. **(c)** The frequency of tumors with mutations in the lung ADC cohort is plotted against the frequency of tumors with mutations in the lung SqCC for the 14 genes significantly mutated only in the Pan-Lung cohort.



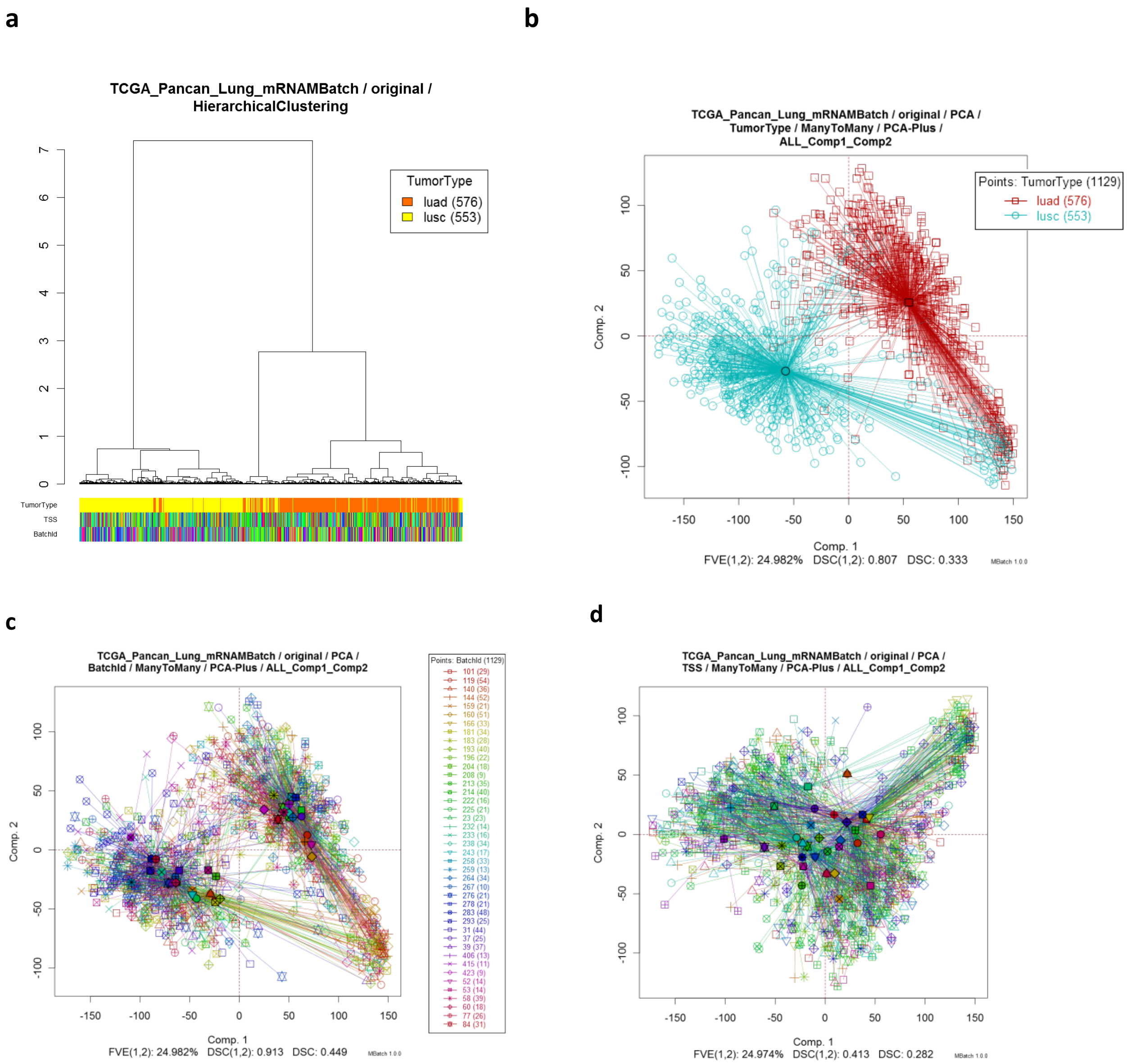
Supplementary Figure 11. Recurrent focal deletions in lung cancer. (a) The *q*-value for deletions in lung ADC are plotted against the best *q*-value for the same gene across 9 other non-lung tumor types[4]. **(b)** The *q*-values for deletions in lung SqCC are compared against 7 other tumor types excluding HNSC and BLCA. Size of the point indicates the frequency of focal alterations. Brackets around gene names indicate that the most likely target gene was inferred from Pan-Cancer copy number analysis across 11 tumor types or from the combined Pan-Lung copy number analysis[4]. Genes located within putative fragile sites are indicated in green.



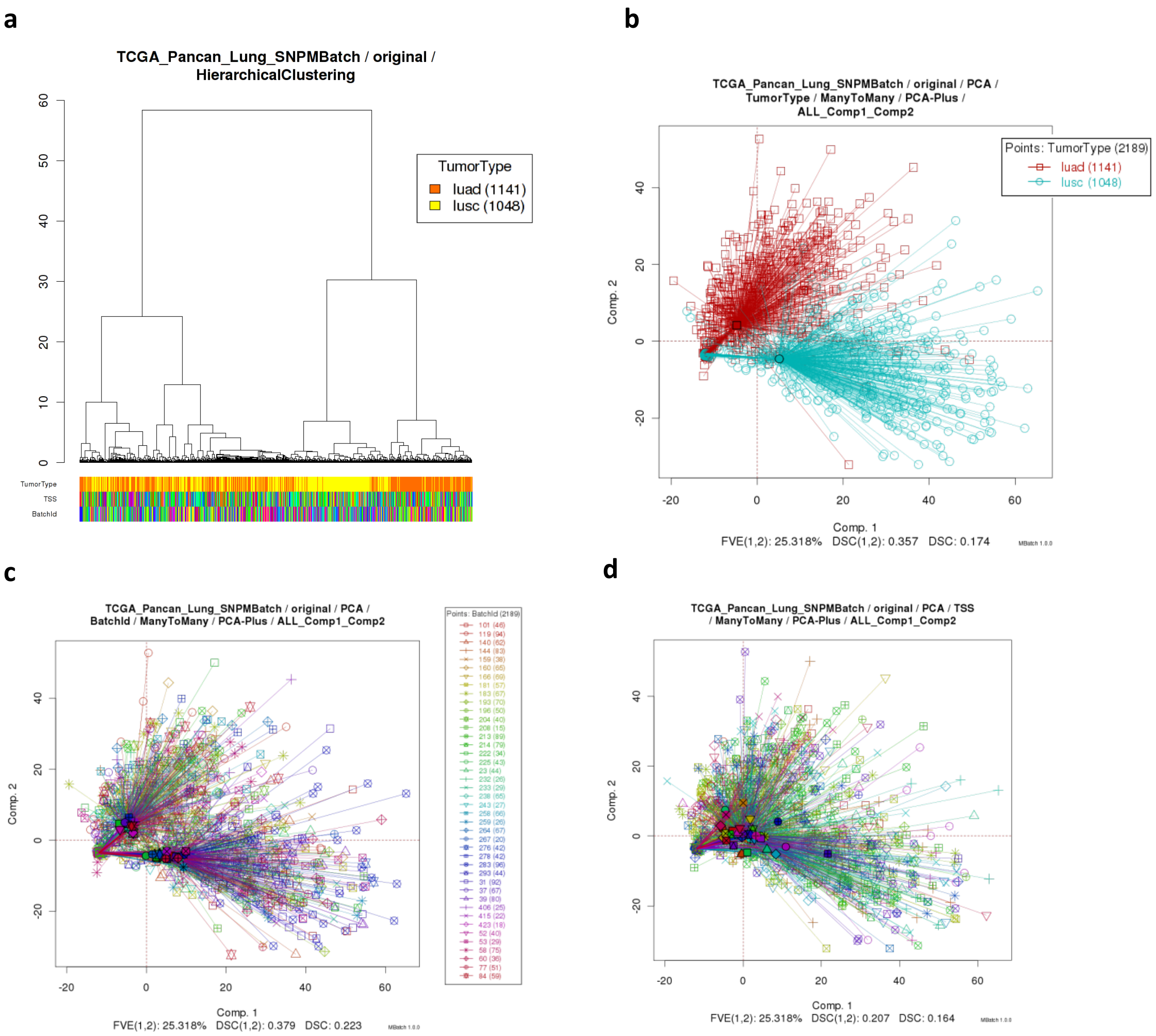
Supplementary Figure 12. Correlation of expression and focal copy number alterations in lung ADC. Threshold copy number estimates for (**a**) focal amplifications or (**b**) focal deletions were obtained from GISTIC2.0 and compared against the $\log_2$ FPKM expression estimates from RNA-sequencing data. Each plot contains a boxplot and a density plot of expression values in each copy number state. A Kruskal-Wallis test was used to associate expression with copy number states.



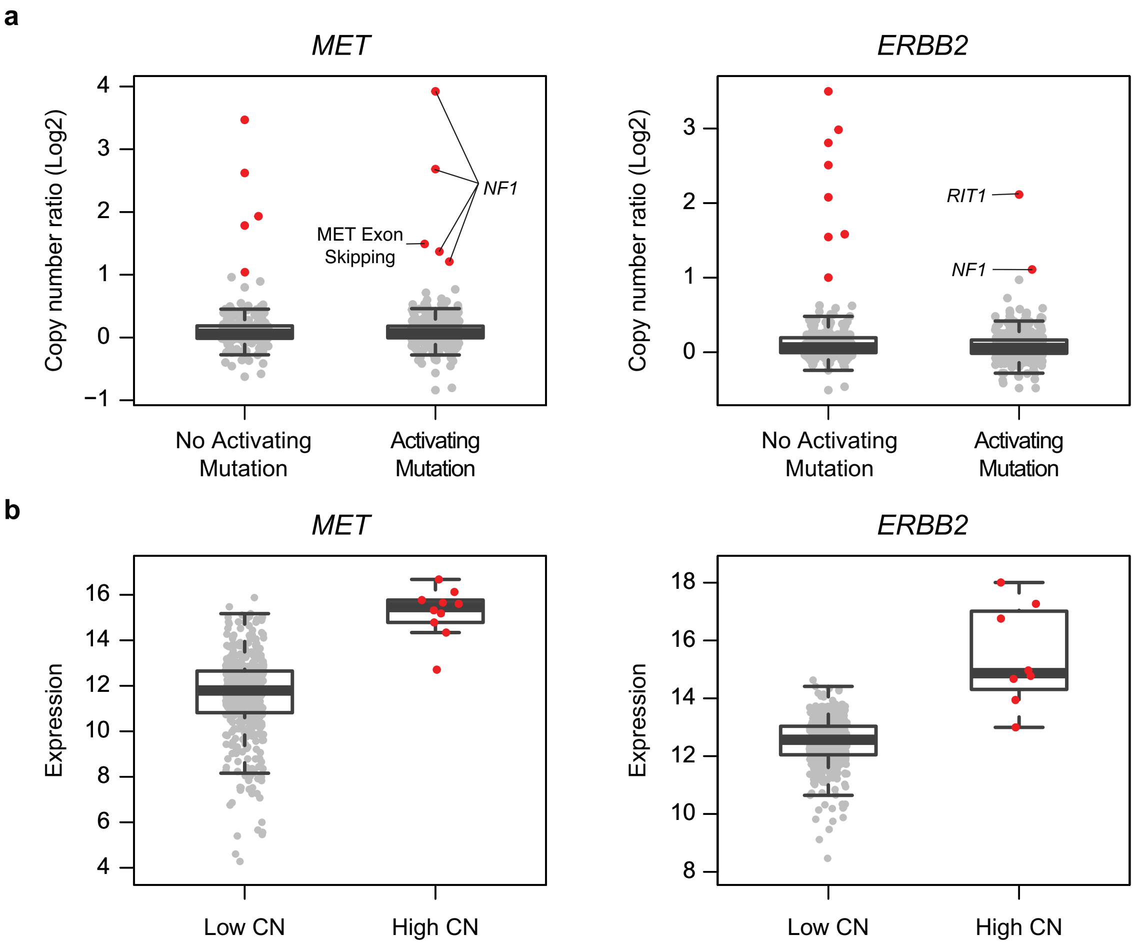
Supplementary Figure 13. Correlation of expression and focal copy number alterations in lung SqCC. Threshold copy number estimates for (**a**) focal amplifications or (**b**) focal deletions were obtained from GISTIC2.0 and compared against the $\log_2$ FPKM expression estimates from RNA-sequencing data. Each plot contains a boxplot and a density plot of expression values in each copy number state. A Kruskal-Wallis test was used to associate expression with copy number states.



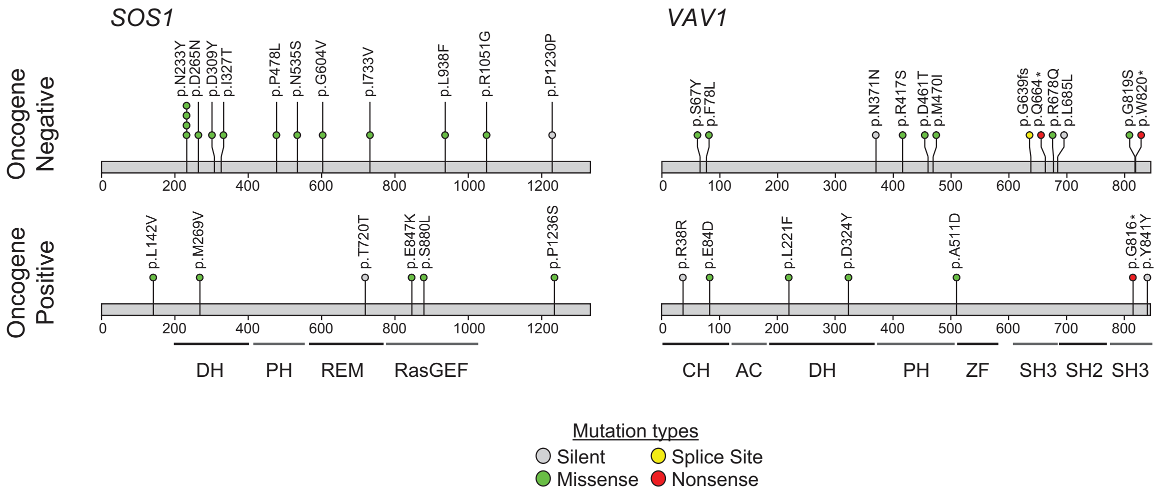
Supplementary Figure 14. Batch effects analysis between lung ADC and lung SqCC mRNA data (sequencing platform). (a) Hierarchical clustering plot for Pan-Lung mRNA data showing covariate bars for tumor type (top), tissue source site (TSS) (middle), and batch ID (bottom). PCA-plot connected by group centroids for tumor type **(b)**, batch ID **(c)**, and tissue source site **(d)** are also shown for the first two principal components. The results show no major clusters by either batch ID or tissue source site, indicating no major batch effects by those two variables. As expected, two large clusters based on tumor type can be seen, along with a third cluster consisting of a mixture of the two tumor types.



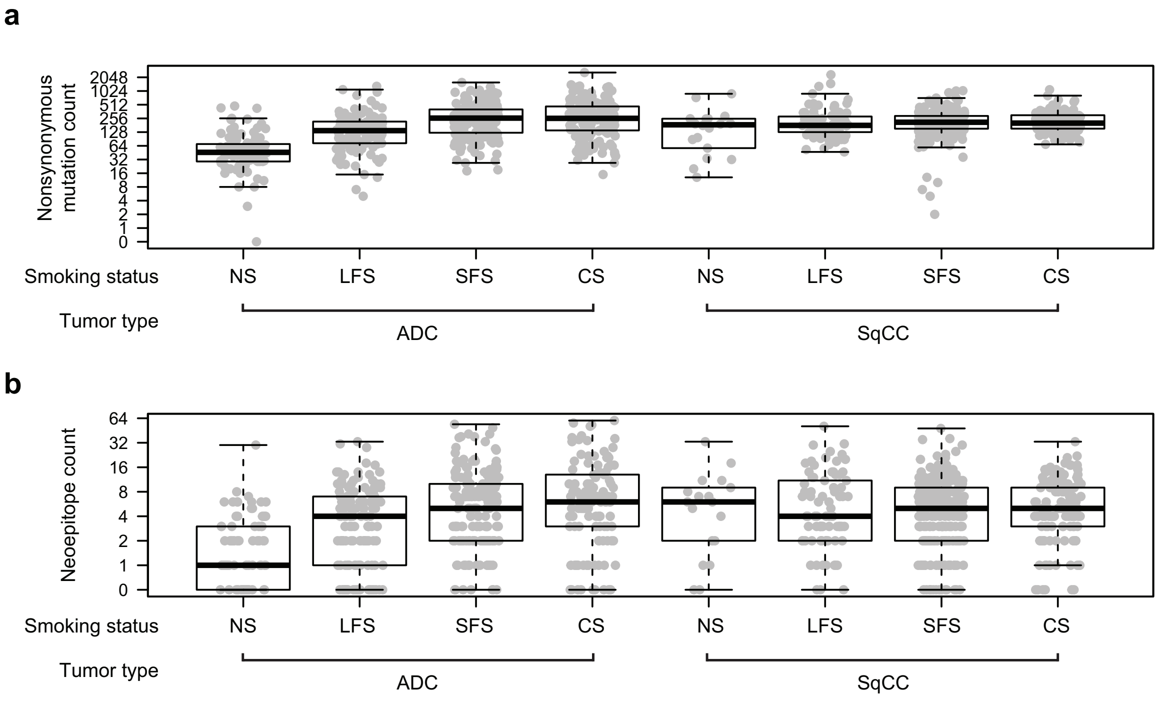
Supplementary Figure 15. Batch effects analysis between lung ADC and lung SqCC CNV data (SNP 6 platform). (a) Hierarchical clustering plot for Pan-Lung CNV data showing covariate bars for tumor type (top), tissue source site (TSS) (middle), and batch ID (bottom). PCA-plot connected by group centroids for tumor type **(b)**, batch ID **(c)**, and tissue source site **(d)** are also shown for the first two principal components. The results show no major clusters by either batch ID or tissue source site, indicating no major batch effects by those two variables. Two clusters based on tumor type can be seen, along with a third cluster consisting of a mixture of the two tumor types. The third cluster contains mostly relatively quiet copy number samples. No major batch effects due to batch ID or TSS were observed.



Supplementary Figure 16. Enrichment of *MET* and *ERBB2* amplification in tumors without other Ras/Raf/RTK mutations. In lung adenocarcinoma, we have previously shown that *MET* and *ERBB2* were recurrently amplified and had high expression in some tumors that lack other activating mutations in the Ras/Raf/RTK pathway[5]. As we did not have RNA-seq data on all samples in the current cohort, we estimated what level of amplification was needed for enrichment in oncogene negative samples. **(a)** If *NF1* and *MET* mutant samples were excluded, tumors with *MET* amplification (total $\log_2$ copy number ratio greater than one) were enriched among tumors without other alterations ($p = 0.005$; Fisher's exact test). Interestingly, four samples with high *MET* amplification also had *NF1* mutations ($p = 0.019$) suggesting a potential synergistic effect. One tumor also exhibited both high *MET* amplification and evidence of *MET* exon 14 skipping. Tumors with an *ERBB2* amplification (total $\log_2$ copy number ratio greater than one) were enriched among tumors without other alterations ($p = 0.002$). One tumor with high *ERBB2* amplification also had a *RIT1* mutation (p.F82L) while another tumor also had an *NF1* mutation. **(b)** In samples with available expression data from RNA-seq, tumors classified as having high levels of *MET* or *ERBB2* amplification (High CN) also displayed higher MET and ERBB2 mRNA expression ($\log_2$), respectively compared to tumors without amplification (Low CN) ($p < 0.001$; Wilcoxon rank-sum test).



Supplementary Figure 17. Mutational patterns for putative Ras/Raf pathway genes in oncogene negative tumors. The frequency of mutations in oncogene negative (n=242) vs. oncogene positive tumors (n=418) was examined for genes that were significantly mutated according to MutSig2CV in either the lung ADC, lung SqCC, or the Pan-Lung cohorts and for genes with known roles in Ras signal transduction[6]. *SOS1* (left) and *VAV1* (right) were significantly enriched in oncogene negative tumors (top) compared to oncogene positive tumors (bottom) (FDR *q*-value < 0.1).



Supplementary Figure 18. Association of nonsynonymous mutation counts and neoepitope counts with smoking status. The immunogenicity of each non-synonymous missense mutation was predicted after inferring HLA alleles within each tumor. **(a)** Overall nonsynonymous mutation counts and **(b)** neoepitope counts were strongly associated with smoking status in lung ADC ($p < 0.001$; Kruskal-Wallis test). These associations were not observed for lung SqCC ($p > 0.05$; Kruskal-Wallis test). Patients included life-long never smokers (NS), longer-term former smokers (LFS) that had quit for at least 15 years, shorter-term former smokers (SFS) that had quit within the last 15 years, and current smokers (CS).

Supplementary references

1. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J. & Stratton, M.R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**, 246-59 (2013).
2. Quesada, V. *et al.* Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet* **44**, 47-52 (2012).
3. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).
4. Zack, T.I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**, 1134-40 (2013).
5. Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-50 (2014).
6. Stephen, A.G., Esposito, D., Bagni, R.K. & McCormick, F. Dragging ras back in the ring. *Cancer Cell* **25**, 272-81 (2014).